# ON ESTIMATING A DYNAMIC FUNCTION OF A STOCHASTIC SYSTEM WITH AVERAGING

LIPTSER, R. AND SPOKOINY, V.

*Dept. Electrical Engineering-Systems,*
*Tel Aviv University,*
*69978 Tel Aviv, Israel*

*and*

*Weierstrass Institute for Applied Analysis and Stochastics,*
*Mohrenstr. 39, 10117 Berlin, Germany*

ABSTRACT. We consider a two-scaled diffusion system, when drift and diffusion parameters of the "slow" component are contaminated by the " fast" unobserved component. The goal is to estimate the dynamic function which is defined by averaging the drift coefficient of the "slow" component w.r.t. the stationary distribution of the "fast" one. We apply a locally linear smoother with a data-driven bandwidth choice. The procedure is fully adaptive and nearly optimal up to a *log log* factor.

# 1. Introduction

In this paper, we propose a procedure for adaptive estimation of "averaged" characteristics of a two-scaled diffusion system described by the Itô equations (w.r.t. independent Wiener processes $w_t$, $W_t$) with a small parameter $\varepsilon$:

$$\mathrm{d}X_t^\varepsilon = f(X_t^\varepsilon, Y_t^\varepsilon)\,\mathrm{d}t + g(X_t^\varepsilon, Y_t^\varepsilon)\,\mathrm{d}w_t, \qquad X_0^\varepsilon = x_0, \tag{1.1}$$

$$\varepsilon\,\mathrm{d}Y_t^\varepsilon = F(Y_t^\varepsilon) + \sqrt{\varepsilon}G(Y_t^\varepsilon)\,\mathrm{d}W_t, \qquad Y_0^\varepsilon = y_0. \tag{1.2}$$

Hereafter, $X_t^\varepsilon$ and $Y_t^\varepsilon$ are referred to as the "slow" and "fast" components respectively. All the functions $f, g, F, G$, entering in (1.1) and (1.2), are unknown and only the slow component $X^\varepsilon$ is observed. The goal is to recover from the observations $X_t^\varepsilon$, $0 \le t \le T$, some characteristics of the process $X^\varepsilon$ which can be used for a further statistical analysis of this process or forecasting.

Examples of such problems meet, for instance, in satellite imaging, where $X_t^\varepsilon$ describes the observed signal and $Y_t^\varepsilon$ is used to describe rotation and vibration of

1991 *Mathematics Subject Classification.* 62G05; Secondary 62M99.

*Key words and phrases.* fast and slow components, drift and diffusion coefficients, averaging principle, nonparametric estimation, bandwidth selection.

The research was carried out within SFB 373 at Humboldt University, Berlin and was printed using funds made available by the Deutsche Forschungsgemeinschaft
The authors are grateful to A.Juditsky and S.Pergamenshchikov for very helpful remarks and discussion .

the satellite. Similar reasoning applies for every measurement device installed on a moving object like plane, train, satellite, car etc. One more reasonable example is connected to asset price processes in financial markets. A weekly (or monthly) observed asset price process $X^\varepsilon$ can be interpreted as the "slow" component. If we are interested in some "global" (macro) characteristics of this process, then the influence of other components of the market can be modeled via the "fast" process $Y^\varepsilon_t$. Some other applications of such approach to the control theory can be found in Kushner (1990) or Liptser, Runggaldier, Taksar (1996).

Equations of the form $\mathrm{d}X_t = f(X_t + Y_t)\,\mathrm{d}t + \mathrm{d}w_t$ are often used to model regression problems with errors in regressors. It is well known, see e.g. Carrol and Hall (1988), Fan and Truong (1993) that the presence of the "error" component $Y_t$ in the regressor variable makes the problem of estimating the regression function $f$ much more difficult. Even if the distribution of $Y_t$ is known, the optimal rate of estimating the function $f$ is only logarithmic in the observation time. We do not assume special additive structure for the arguments of the drift function $f$ and no information about the distribution of the noisy component $Y$ is available. Instead we only assume that $Y^\varepsilon$ is a fast oscillating process. We shall see that this qualitative assumption allows for a reasonable quality of estimation of the "averaged" drift function $\overline{f}$ which describes the "macro" characteristics of the process $X^\varepsilon$.

It is well known from Khasminskii (1966) (see also Freidlin and Wentzell (1984), Veretennikov (1991)) that, under some regularity conditions on the functions $F$ and $G$ from (1.2), $Y^\varepsilon$ is a fast oscillating ergodic process while the slow process $X^\varepsilon$ obeys, so called, Bogolubov's averaging principle. This roughly means that the distribution of the slow component is close to the distribution of the diffusion process $(X_t)_{t\ge 0}$ defined by the Itô equation

$$\mathrm{d}X_t = \overline{f}(X_t)\,\mathrm{d}t + \overline{g}(X_t)\,\mathrm{d}\overline{w}_t, \tag{1.3}$$

where $\overline{w}$ is some Wiener process and the drift and diffusion coefficients $\overline{f}, \overline{g}$ are defined by averaging the original coefficients with respect to the stationary density $p$ of the fast process:

$$\overline{f}(x) = \int f(x,y)\,p(y)\,\mathrm{d}y \qquad \text{and} \qquad \overline{g}(x) = \left(\int g^2(x,y)\,p(y)\,\mathrm{d}y\right)^{1/2}.$$

In other words, the "macro" behavior of the process $X^\varepsilon$ is determined only by the averaged functions $\overline{f}$ and $\overline{g}$. This naturally leads to the problem of statistical estimation of these functions from observations $X^\varepsilon_t$, $0 \le t \le T$, where $T$ is the *observation time*.

In this paper, we focus on estimating the *dynamic function* $\overline{f}(x)$. We do not discuss here the problem of estimating the diffusion coefficient $\overline{g}$ since in the case of continuous observations, the function $\overline{g}$ can be estimated at an essentially better rate (of order $\varepsilon^2$) than the drift $\overline{f}$. We also restrict ourselves to the problem of pointwise estimation, that is, given a point $x$, we estimate the value $\overline{f}(x)$. We refer to Lepski, Mammen and Spokoiny (1997) for a discussion of the relation between pointwise and global estimation. Note that the problem of the pointwise estimation of the drift function $f$ is closely connected to the problem of forecasting the process $X^\varepsilon$. Indeed, if we observe the process $(X^\varepsilon_t)$ until the time-point $T$, and if we are interested in a behavior of the process in the nearest future after $T$, then we have to estimate $\overline{f}(x)$ for $x = X_t^\varepsilon$.

The estimation theory for diffusion type processes is well developed under the parametric modeling when underlying functions (drift and diffusion) are specified up to a value of a finite dimensional parameter (cf. Kutoyants, 1984b). In contrast, nonparametric estimation is not studied in details. The known results concern only with statistical inference for ergodic diffusion models with a small noise or for a large observation time $T$. Kutoyants (1984a) evaluated the minimax rate of estimation of the drift coefficient using a kernel type estimator. Genon-Catalot, Laredo and Picard (1992) applied wavelets. Locally polynomial estimators are described in Fan and Gijbels (1996). Milstein and Nussbaum (1994), Grama and Nussbaum (1998) established the LeCam equivalence between the diffusion model and the "white noise model". Some pertinent results for autoregressive models in discrete time can be found in Doukhan and Ghindes (1980), Collomb and Doukhan (1983), Doukhan and Tsybakov (1993), Delyon and Juditsky (1997).

In this paper, we assume neither ergodic properties of the slow component nor the large observation time $T$. This makes the problem more complicated. Additional difficulties come from the fact that the coefficients of the slow process are contaminated by the unobserved fast one. To our knowledge, nonparametric statistical inference for diffusion models (1.1), (1.2) with averaging has not yet been considered.

We propose a locally linear estimator of $\overline{f}(x)$ with a data-driven bandwidth choice and show that this method provides a nearly optimal rate of estimation up to a *log log* factor.

The paper is organized as follows. The next section contains the description of the locally linear estimator. Its properties are discussed in Section 3. The data-driven bandwidth choice is presented in Section 4. All proofs are gathered in Sections 5.

# 2. A locally linear estimator

For fixed $x$, to estimate the value $\overline{f}(x)$ we apply the locally linear smoother (cf. Katkovnik (1985), Tsybakov (1986), Fan and Gijbels (1996)).

We begin with some heuristic explanations of the method. Imagine for a moment that the observed process $X_t$, $0 \le t \le T$ satisfies the Itô equation with respect to Wiener process $w_t$:

$$\mathrm{d}X_t = f(X_t)\,\mathrm{d}t + g(X_t)\,\mathrm{d}w_t \tag{2.1}$$

with the linear function $f$: $f(u) = \theta_0 + \theta_1 \dfrac{u-x}{h}$, depending on two parameters $\theta_0, \theta_1$, where $x$ and $h > 0$ are fixed. These parameters can be estimated by the least squares method:

$$(\widetilde{\theta}_0, \widetilde{\theta}_1) = \underset{\theta_0,\theta_1}{\operatorname{argmax}} \left\{ \int_0^T \left(\theta_0 + \theta_1 \frac{X_t - x}{h}\right) \mathrm{d}X_t - \frac{1}{2}\int_0^T \left(\theta_0 + \theta_1 \frac{X_t - x}{h}\right)^2 \mathrm{d}t \right\},$$

that is, with $\mu_k = \int_0^T \left(\frac{X_t - x}{h}\right)^k \, \mathrm{d}t$, $k = 0, 1, 2$, we get

$$\begin{aligned}
\widetilde{\theta}_0 &= \frac{\mu_2 \int\limits_0^T \mathrm{d}X_t - \mu_1 \int\limits_0^T \frac{X_t - x}{h} \, \mathrm{d}X_t}{\mu_0 \mu_2 - \mu_1^2}, \\
\widetilde{\theta}_1 &= \frac{-\mu_1 \int\limits_0^T \mathrm{d}X_t + \mu_0 \int\limits_0^T \frac{X_t - x}{h} \, \mathrm{d}X_t}{\mu_0 \mu_2 - \mu_1^2}.
\end{aligned}$$

Since clearly $f(x) = \theta_0$, the value $\widetilde{\theta}_0$ can be taken as the estimate of $f(x)$.

The locally linear smoother is defined in a similar way. The only difference is that the function $f$ is not assumed to be linear but it is approximated by a linear function $\theta_0 + \theta_1 \dfrac{u - x}{h}$ in a small neighborhood $[x - h, x + h]$ of the point $x$. Then the coefficients $\theta_0, \theta_1$ of this function can be estimated from the observations $X_t$ falling into the interval $[x - h, x + h]$, $0 \leq t \leq T$. For the formal description, let us introduce the *kernel* function $Q(u)$ which is assumed to be smooth, non-negative, bounded by 1, and vanishing outside of $[-1, 1]$. Then the locally linear estimate with the kernel $Q$ and a *bandwidth* $h$ is defined as:

$$\widetilde{f}_h(x) = \frac{\mu_{2,h} \int\limits_0^T Q\left(\frac{X_t - x}{h}\right) \mathrm{d}X_t - \mu_{1,h} \int\limits_0^T \frac{X_t - x}{h} Q\left(\frac{X_t - x}{h}\right) \mathrm{d}X_t}{\mu_{0,h} \mu_{2,h} - \mu_{1,h}^2}, \tag{2.2}$$

where

$$\mu_{k,h} = \int_0^T \left(\frac{X_t - x}{h}\right)^k Q\left(\frac{X_t - x}{h}\right) \mathrm{d}t, \qquad k = 0, 1, 2.$$

Now we come back to the more complicated two-scaled model (1.1), (1.2). Here, due to the averaging principle, the observed process $X_t^\varepsilon$ is closed in the distribution sense to the "limit" process $X_t$ described by equation (1.3). Therefore, to define our estimate $\widetilde{f}_h(x)$ of $\overline{f}(x)$, we simply replace in expression (2.2) the "limit" process $X_t$ by our observations $X_t^\varepsilon$:

$$\widetilde{f}_h(x) = \frac{\mu_{2,h} \int\limits_0^T Q\left(\frac{X_t^\varepsilon - x}{h}\right) \mathrm{d}X_t^\varepsilon - \mu_{1,h} \int\limits_0^T \frac{X_t^\varepsilon - x}{h} Q\left(\frac{X_t^\varepsilon - x}{h}\right) \mathrm{d}X_t^\varepsilon}{\mu_{0,h} \mu_{2,h} - \mu_{1,h}^2}, \tag{2.3}$$

where now

$$\mu_{k,h} = \int_0^T \left(\frac{X_t^\varepsilon - x}{h}\right)^k Q\left(\frac{X_t^\varepsilon - x}{h}\right) \mathrm{d}t, \qquad k = 0, 1, 2. \tag{2.4}$$

The quality of estimate (2.3) essentially depends on the bandwidth $h$. Some useful properties of $\widetilde{f}_h(x)$ for the fixed $h$ are described in Section 3. We discuss the adaptive choice of the bandwidth $h$ in Section 4.

# 3. Accuracy of the locally linear estimate

In this section we study some properties of the locally linear estimate $\widetilde{f}_h(x)$ from (2.3). We first formulate the required conditions on the coefficients of the two-scaled system (1.1), (1.2). Then we present the result and discuss some its corollaries.

## 3.1. Conditions

In the sequel we suppose that the functions $f, g$ and $F, G$ from (1.1) and (1.2) obey the following conditions:

$(A_s)$ Functions $f(x, y)$ and $g(x, y)$ are Lipschitz continuous in $x, y$ and $f(x, y)$ is three times continuously differentiable in $x$. For some positive constants $g_{\min} \le g_{\max}$

$$g_{\min} \le |g(x, y)| \le g_{\max}.$$

$(A_f)$ 1. Functions $F(y)$ and $G(y)$ are Lipschitz continuous in $y$ and continuously differentiable ( $F$ once, $G$ twice) and their derivatives are continuous and bounded.

2. There exist constants $\kappa > 0$ and $C > 1$ such that for $|y| > C$

$$yF(y) \le -\kappa |y|^2,$$

3. Function $G$ is bounded and strongly positive, i.e. for any $y$

$$0 < G_{\min} \le |G(y)| \le G_{\max}.$$

Condition $(A_f)$ guarantees the required ergodicity of the fast process $Y_t^{\varepsilon}$ and, moreover, this condition can be viewed as the mathematical formulation of the ergodic property of the fast process [1]. Under $(A_f)$ the invariant density of the fast process can be explicitly described (Khasminskii, 1966) and it does not depend on $\varepsilon$:

$$p(y) = \text{Const.} \frac{\exp\left\{ 2 \int\limits_0^y \frac{F(u)}{G^2(u)} \, \mathrm{d}u \right\}}{G^2(y)}. \tag{3.1}$$

It is worth to mention that neither the constants $C, \kappa, G_{\min}, G_{\max}$, nor the invariant density $p$ are not assumed to be known and they do not enter into the description of the procedure and into the formulation of the main results.

[1] see Veretennikov (1991) for more detailed analysis of $(A_f)$

## 3.2. Accuracy of the locally linear estimate

To state the result, we introduce some additional notations. With $\mu_{k,h}$ defined in (2.4), set $D_h = \mu_{0,h}\mu_{2,h} - \mu_{1,h}^2$ , and

$$
\begin{aligned}
(3.2)\quad \sigma_h^2(x) &= \frac{1}{D_h^2}\int_0^T \left(\mu_{2,h} - \mu_{1,h}\frac{X_t^\varepsilon - x}{h}\right)^2 Q^2\left(\frac{X_t^\varepsilon - x}{h}\right) g^2(X_t^\varepsilon, Y_t^\varepsilon)\,\mathrm{d}t \\
&= v_{2,h}^2 V_{0,h} - 2v_{1,h}v_{2,h}V_{1,h} + v_{1,h}^2 V_{2,h}
\end{aligned}
$$

where, with $k = 0, 1, 2$ ,

$$
(3.3)\qquad v_{k,h} = \frac{\mu_{k,h}}{D_h} = \frac{\mu_{k,h}}{\mu_{0,h}\mu_{2,h} - \mu_{1,h}^2},
$$

$$
(3.4)\qquad V_{k,h} = \int_0^T \left(\frac{X_t^\varepsilon - x}{h}\right)^k Q^2\left(\frac{X_t^\varepsilon - x}{h}\right) g^2(X_t^\varepsilon, Y_t^\varepsilon)\,\mathrm{d}t.
$$

Although the expressions for $V_{k,h}$ , $k = 0, 1, 2$ , use the unknown diffusion coefficient $g(X_t^\varepsilon, Y_t^\varepsilon)$ and moreover, one of its arguments $Y_t^\varepsilon$ is not observed, these values can be computed on the base of our observations $(X_t^\varepsilon,\ 0 \le t \le T)$ only, see Section 3.4.

The value $\sigma_h^2(x)$ is called the *conditional variance* of the estimate $\widetilde{f}_h(x)$ . We use this terminology by analogy with the regression case, where $X_t^\varepsilon$ is a deterministic design process and $\sigma_h^2(x)$ is really the variance of the least squares estimate $\widetilde{f}_h(x)$ . Note that for the regression setup, some design regularity is required to ensure that $\sigma_h^2(x)$ is not too large.

In our case, the observed process $(X_t^\varepsilon)_{t\ge 0}$ is described by the autoregressive type equation and it can be viewed at the same time as the design process. We therefore impose some conditions on the trajectories of the process $X_t^\varepsilon$ which are similar to that of used to describe the design regularity in the regression setting. Our results are also similar to that of usually obtained in the regression estimation. In particular, we show that under the conditions imposed, the conditional variance $\sigma_h^2(x)$ helps to control the stochastic component of the estimate $\widetilde{f}_h(x)$ .

For some $\rho \ge 0$ , $r > 0$ , $b > 0$ and $B \ge 1$ , we introduce the set

$$
\mathcal{A}_h = \left\{
\begin{array}{ll}
b \le Th\, v_{2,h} \le bB\,, & b \le Th\,\sigma_h^2(x) \le bB\,, \\
\mu_{0,h} \le r\mu_{2,h}\,, & V_{0,h} \le rV_{2,h} \\
\mu_{1,h}^2 \le \rho\mu_{0,h}\mu_{2,h}\,, & V_{1,h}^2 \le \rho V_{0,h}V_{2,h}
\end{array}
\right\}.
$$

Since $(X_t^\varepsilon)_{t\ge 0}$ is the random process, the set $\mathcal{A}_h$ is random as well. In the sequel we study the properties of $\widetilde{f}_h(x)$ restricted to the set $\mathcal{A}_h$ , see Remark 3.1 and 3.2 for further discussion.

The quality of the approximation of $f(u, y)$ by a linear in $u$ function in the neighborhood $u \in [x - h, x + h]$ is characterized by the following quantity

$$
(3.5)\qquad \Delta_h(x) = \sup_{|u-x|\le h,\, y\in\mathbb{R}} |f(u, y) - f(x, y) - (u - x)f_x(x, y)|.
$$

The next theorem describes some useful properties of estimate (2.3).

**Theorem 3.1.** *Let $(A_s)$ and $(A_f)$ be fulfilled, and let the values $\varepsilon$ and $\varepsilon T$ be sufficiently small and $Th \geq 1$. Then for every $\lambda \geq \sqrt{2}$*

$$\boldsymbol{P}\left(\left|\widetilde{f}_h(x) - \overline{f}(x)\right| > c\Delta_h(x) + \lambda\sigma_h(x),\, \mathcal{A}_h\right) \leq 4e\log(4B^3)\left(1 + 4r\sqrt{\frac{1+r}{1-\rho}}\,\lambda^2\right)\lambda e^{-\frac{\lambda^2}{2}}, \tag{3.6}$$

*where* $c = (1-\rho)^{-1/2}$.

*Remark* 3.1. As we mentioned previously, the quality of the estimate $\widetilde{f}_h(x)$ is examined on the set $\mathcal{A}_h$ only. This allows to eliminate irregular cases when, for instance, the trajectory $X^\varepsilon_{[0,T]}$ does not pass through the interval $[x-h, x+h]$ and hence $\mu_{0,h} = \mu_{1,h} = \mu_{2,h} = D_h = 0$. Note that for typical applications to forecasting, when we have to estimate $\overline{f}(x)$ with $x = X^\varepsilon_t$, the trajectory $X^\varepsilon_{[0,T]}$ obviously passes through $x$.

Note also that the event $\mathcal{A}_h$ is completely determined by the known values $\mu_{k,h}$ and $V_{k,h}$, $k = 0,1,2$. We therefore always know whether the observed trajectory $X^\varepsilon_{[0,T]}$ belongs to $\mathcal{A}_h$ or not. If the trajectory $X^\varepsilon_{[0,T]}$ does not belong to $\mathcal{A}_h$, we are not able to guarantee a reasonable quality of estimation for $\widetilde{f}_h(x)$.

The conditions $0 \leq Q(u) \leq 1$ and $Q(u) = 0$ for $|u| \geq 1$ imply $\mu_{2,h} \leq \mu_{0,h}$ and $V_{2,h} \leq V_{0,h}$. Further, by the Cauchy-Schwarz inequality, it holds $\mu^2_{1,h} \leq \mu_{0,h}\mu_{2,h}$ and $V^2_{1,h} \leq V_{0,h}V_{2,h}$. The conditions $\mu_{0,h} \leq r\mu_{2,h}$, $V_{0,h} \leq rV_{2,h}$, $\mu^2_{1,h} \leq \rho\mu_{0,h}\mu_{2,h}$ and $V^2_{1,h} \leq \rho V_{0,h}V_{2,h}$ with $\rho < 1$ and $r \geq 1$ ensure that the local linear estimate is well defined. Note that these conditions are not completely independent. In particular, if $g(x,y)$ is a constant function and if $Q(u) = \mathbf{1}(|u| \leq 1)$, then $\mu_{k,h} = V_{k,h}$, $k = 0,1,2$ and $\sigma^2_h(x) = v_{2,h} = \mu_{2,h}/(\mu_{0,h}\mu_{2,h} - \mu^2_{1,h})$.

The choice of constants $\rho$, $b$, $B$, $r$, entering in the definition of the set $\mathcal{A}_h$, is optional and they even may depend on $\varepsilon$ and $T$. Note that the upper bound (3.6) from Theorem 3.1 does not depends on $b$ and it depends on $B$ (which determines the variability of the conditional variance $\sigma^2_h(x)$) only via the log-factor $\log(4B^3)$.

*Remark* 3.2. If the coefficients $f$ and $g$ of the slow component obey conditions similar to $(A_f)$, then $(X^\varepsilon_t)_{t\geq 0}$ is also the ergodic process and its transition probabilities converge to some stationary distribution as the observation time tends to infinity, see e.g. Veretennikov (1991). This particularly means that the normalized integrals $(Th)^{-1}\mu_{k,h}$ and $(Th)^{-1}V_{k,h}$ ($k = 0,1,2$) converge to some fixed values which depend only on the stationary distribution of $(X^\varepsilon_t)$. Therefore we can select fixed constants $b, B$ and $\rho, r$ in such a way that $1 - \boldsymbol{P}(\mathcal{A}_h)$ converges to zero exponentially fast as $T \to \infty$. Since obviously

$$\begin{aligned}\boldsymbol{P}&\left(\left|\widetilde{f}_h(x) - \overline{f}(x)\right| > c\Delta_h(x) + \lambda\sigma_h(x)\right)\\ &\leq \boldsymbol{P}\left(\left|\widetilde{f}_h(x) - \overline{f}(x)\right| > c\Delta_h(x) + \lambda\sigma_h(x),\, \mathcal{A}_h\right) + \boldsymbol{P}(\mathcal{A}_h)\end{aligned}$$

we obtain in this situation an unconditional asymptotic bound for the risk of the estimate $\widetilde{f}_h(x)$.

*Remark* 3.3. The result of the theorem claims that the losses $|\widetilde{f}_h(x) - \overline{f}(x)|$ of the estimate $\widetilde{f}_h(x)$, being restricted to $\mathcal{A}_h$, are bounded by the sum of two terms: $c\Delta_h(x)$ and $\lambda\sigma_h(x)$. The first one mimics the accuracy of approximating the function $f(u,y)$ by a linear in $u$ function in the small vicinity $[x-h, x+h]$ of $x$. The second term is in proportion to the "stochastic standard deviation" $\sigma_h(x)$. Note also that the definition of the set $\mathcal{A}_h$ provides $\sigma_h(x) \asymp (Th)^{-1/2}$, where the symbol "$\asymp$" means equivalence in order.

## 3.3. Quality of estimation under smoothness assumptions

Due to the assumptions $(A_s)$ from Section 3, the function $f$ is twice continuously differentiable with respect to the first argument. Assume also that for every $u$ from a small vicinity of $x$ and any fixed $y$

$$\left| \frac{\partial^2 f(u,y)}{\partial u^2} \right| \le L. \tag{3.7}$$

Then the value $\Delta_h(x)$ defined in (3.5), is bounded above by $Lh^2/2$. On the other hand, on the set $\mathcal{A}_h$ the stochastic variance $\sigma_h^2(x)$ is of order $(Th)^{-1}$. Therefore, following to the standard approach in nonparametric estimation, the bandwidth $h$ can be chosen by balancing the accuracy of approximation and the stochastic error:

$$Lh^2 \asymp \frac{1}{\sqrt{T\,h}}.$$

This leads to the choice $h \asymp (T\,L^2)^{-1/5}$ and hence to the rate of the estimation $L^{1/5}T^{-2/5}$ which is optimal in the minimax sense under the smoothness assumptions (3.7), see e.g. Ibragimov and Khasmiskii (1981). Unfortunately this approach hardly applies in practice, since the constant $L$ in (3.7) is typically unknown. An adaptive (data-driven) choice of the bandwidth is discussed in the next section.

## 3.4. Computation of $\sigma_h^2(x)$

Recall that with fixed $h$, the value $\sigma_h^2(x)$ is defined by the formulas (3.2) through (3.4) where the expressions for $V_{k,h}$, $k = 0,1,2$, use the unknown diffusion coefficient $g^2(X_t^\varepsilon, Y_t^\varepsilon)$ and the unobserved process $Y_t^\varepsilon$ as one of its arguments. We now show that despite of this fact, the value $\sigma_h^2(x)$ can be computed via the trajectory $X_{[0,T]}^\varepsilon$ only.

Let us introduce two random processes

$$Z_t' = \int_0^t Q\left(\frac{X_s^\varepsilon - x}{h}\right) \mathrm{d}X_s^\varepsilon \quad \text{and} \quad Z_t'' = \int_0^t Q\left(\frac{X_s^\varepsilon - x}{h}\right) \frac{X_s^\varepsilon - x}{h} \,\mathrm{d}X_s^\varepsilon$$

which are completely determined on the time interval $[0,T]$ by the observation $X^{\varepsilon}_{[0,T]}$. Applying the Itô formula we get

$$\begin{aligned}
(Z'_T)^2 &= 2\int_0^T Z'_t \,\mathrm{d}Z'_t + V_{0,h} \\
(Z''_T)^2 &= 2\int_0^T Z''_t \,\mathrm{d}Z''_t + V_{2,h} \\
Z'_T Z''_T &= \int_0^T Z'_t \,\mathrm{d}Z''_t + \int_0^T Z''_t \,\mathrm{d}Z'_t + V_{1,h}.
\end{aligned}$$

Hence $V_{0,h} = (Z'{}_T)^2 - 2\int\limits_0^T Z'_t \,\mathrm{d}Z'_t$, so that $V_{0,h}$ is completely determined by $X^{\varepsilon}_{[0,T]}$. Similar arguments apply for $V_{1,h}$ and $V_{2,h}$ and hence for $\sigma_h^2(x)$ as required.

# 4. Data-driven bandwidth selection

In this section we consider the problem of bandwidth selection for the locally linear estimator described in Section 2. It is assumed here that the method of estimation, that is the locally linear smoother with the kernel $Q$, is fixed and only the bandwidth $h$ has to be chosen. The adaptive procedure originates from Lepski (1990), see also Lepski, Mammen and Spokoiny (1997) and Lepski and Spokoiny (1997).

## 4.1. An "ideal" bandwidth

First we introduce the notion of an "ideal" bandwidth. Let a set $\mathcal{H}$, of all admissible bandwidths $h$, be fixed. For technical reasons, we assume that this set is finite and denote by $\#\mathcal{H}$ the number of its elements. Usually $\mathcal{H}$ is taken as a geometric grid of the form

$$\mathcal{H} = \{h = h_{\min} a^k,\ k = 0,1,2,\ldots : h \le h_{\max}\},$$

where $h_{\min} \le h_{\max}$ and $a > 1$ are some prescribed constants. As in Section 3, we restrict ourselves only to those $h$ from $\mathcal{H}$ for which the observed trajectory $X^{\varepsilon}_{[0,T]}$ belongs to $\mathcal{A}_h$. Our goal is to select $h$ from $\mathcal{H}$ providing the minimal in some sense error of estimation for the corresponding estimate $\widetilde{f}_h(x)$.

We begin with some heuristic explanations. Recall first, that the values $\sigma_h(x)$ can be exactly computed on the base of observations $X^{\varepsilon}_{[0,T]}$, see Subsection 3.4. Note also that $\sigma_h(x)$ typically decreases in $h$. Indeed, an increase of $h$ makes the estimation window $[x-h, x+h]$ larger and hence more observations can be used for estimating the underlying function $\overline{f}$ at the point $x$. This results in a smaller variance of the estimate. To simplify the exposition, we suppose that $\sigma_h(x)$ strongly decreases in $h \in \mathcal{H}$. (If this assumption is not fulfilled for the original set $\mathcal{H}$, i.e. if there is $h' < h \in \mathcal{H}$ with the property $\sigma_h(x) \ge \sigma_{h'}(x)$, then we simply exclude $h$ from $\mathcal{H}$.)

The behavior of the bias term $\Delta_h(x)$ is just opposite. Namely, for a regular function $f$, the value $\Delta_h(x)$ is small when $h$ is small, and it typically increases in $h$. Therefore, the minimization of the sum of the form $c\Delta_h(x) + \lambda\sigma_h(x)$ with some constants $c, \lambda$ leads to the balance relation $\Delta_h(x) \asymp \sigma_h(x)$ and we define a "good" bandwidth $h^*$ as the largest $h$ from $\mathcal{H}$ such that $c\Delta_h(x)$ is still not larger than $D\sigma_h(x)$ with some prescribed constant $D$:

$$h^* = \max\{h \in \mathcal{H} : c\Delta_h(x) \le D\sigma_h(x)\}. \tag{4.1}$$

Since $\Delta_h(x)$ is unknown, the bandwidth $h^*$ is unknown as well. In the sequel, following to Donoho and Johnstone (1995), $h^*$ is referred to as the "ideal" bandwidth or "oracle". Due to Theorem 3.1, the losses of the "ideal" estimate $\widetilde{f}_{h^*}$ are bounded (with probability closed to one) by $(D+\lambda)\sigma_{h^*}(x)$ provided that $\lambda$ is sufficiently large.

## 4.2. An adaptive bandwidth choice

Now we present our adaptive procedure and show that the corresponding accuracy of the estimation is essentially the same as if the "ideal" bandwidth applies. The procedure involves two positive parameters $\lambda_1$ and $D$. The last one is already mentioned in the definition of the "ideal" bandwidth. We discuss the choice of $\lambda_1$ and $D$ at the end of this section.

The data-driven bandwidth $\widehat{h}$ is defined by the following rule:

$$\widehat{h} = \max\Big\{h \in \mathcal{H} : \big|\widetilde{f}_h(x) - \widetilde{f}_\eta(x)\big| \le \lambda_1\Big(\sigma_h(x) + \sigma_\eta(x)\Big) + 2D\sigma_h(x), \quad \forall \eta \in \mathcal{H}, \eta < h\Big\}. \tag{4.2}$$

In words, the rule prescribes to take the largest value $h \in \mathcal{H}$ for which the corresponding estimate $\widetilde{f}_h(x)$ does not differ essentially from every estimate $\widetilde{f}_\eta(x)$ with a smaller bandwidth value $\eta \in \mathcal{H}$. The arguments for this choice are quite simple: if both $\eta$ and $h$ are not larger than $h^*$, then the "bias" terms $\Delta_\eta(x)$ and $\Delta_h(x)$ in the difference $\big|\widetilde{f}_h(x) - \widetilde{f}_\eta(x)\big|$ are bounded by $2D\sigma_{h^*}(x) \le 2D\sigma_h(x)$ and therefore, the probability of the event

$$\Big\{\big|\widetilde{f}_h(x) - \widetilde{f}_\eta(x)\big| > \lambda_1\Big(\sigma_h(x) + \sigma_\eta(x)\Big) + 2D\sigma_h(x)\Big\}$$

is small provided that $\lambda_1$ is large enough (see Theorem 3.1). Hence, if we meet the opposite inequality for some $\eta < h$, this means that the bias $\Delta_h(x)$ is already too large and the bandwidth $h$ is not a good one.

Finally, to define our adaptive estimate, we plug the data-driven bandwidth $\widehat{h}$ in the estimate $\widetilde{f}_h(x)$:

$$\widehat{f}(x) \equiv \widetilde{f}_{\widehat{h}}(x).$$

In the next theorem we describe some properties of the adaptive estimate $\widehat{f}(x)$ restricted to the set

$$\mathcal{A}^* = \bigcap_{h \in \mathcal{H}} \mathcal{A}_h.$$

**Theorem 4.1.** *Let the values $\varepsilon$ and $\varepsilon T$ be sufficiently small. Let also $h^*$ be defined in (4.1) with $\lambda_1 \ge \sqrt{2}$. Then the estimate $\widehat{f}(x)$ fulfills the following property: for any $\lambda$ with $\sqrt{2} \le \lambda \le \lambda_1$*

$$\boldsymbol{P}\left(\left|\widehat{f}(x)-\overline{f}(x)\right|>(\lambda+2\lambda_1+3D)\sigma_{h^*}(x),\mathcal{A}^*\right) \le 4e\log(4B^3)\left(1+4r\sqrt{\frac{1+r}{1-\rho}}\,\lambda_1^2\right)\lambda_1\left\{(\#\mathcal{H})^2e^{-\frac{\lambda_1^2}{2}}+e^{-\frac{\lambda^2}{2}}\right\}. \tag{4.3}$$

*Remark* 4.1. The choice of parameters $\lambda_1$, $D$, entering in (4.2), plays the important role. The bound in (4.3) shows that the probability for $\left|\widehat{f}(x)-\overline{f}(x)\right|$ of being large is small, provided that the value $(\#\mathcal{H})^2\lambda_1^2e^{-\lambda_1^2/2}$ is sufficiently small. This leads to the choice

$$\lambda_1 \approx \sqrt{4\log(\#\mathcal{H})+\lambda^2}$$

so that

$$(\#\mathcal{H})^2\lambda_1e^{-\lambda_1^2/2}\approx e^{-\lambda^2/2}.$$

If $\mathcal{H}$ is taken in the form of the geometric grid, then we get $\#\mathcal{H}\approx\log_a(h_{\max}/h_{\min})$. Therefore, taking $h_{\max}\approx T$ and $h_{\min}\approx 1$, we arrive at

$$\lambda_1\approx\sqrt{4\log\log T+\lambda^2}.$$

There is much more degree of freedom in the choice of $D$. This parameter controls the balance between the accuracy of approximating the function $f$ by a linear one and the stochastic error (see the definition (4.1) of the "ideal" bandwidth $h^*$). The results from Lepski and Spokoiny (1997) lead to the choice $D=\mathrm{Const}\,\lambda_1$ (see also the next subsection). At the same time, Lepski and Levit (1997) argued that for a smooth function $f$, the relevant choice is $D=0$. Simulation results show a reasonable performance of the presented procedure with $\lambda_1\approx 3$ and $D=0$.

## 4.3. The rate of adaptive estimation

We now compare the accuracy of the adaptive procedure (4.2) with the "optimal" one designed for the case of known smoothness properties of the underlying function $f$ (see Subsection 3.3).

Assume (3.7). Then $\Delta_h(x)\le Lh^2/2$ and the constraints $c\Delta_h(x)\le D\sigma_h(x)$ and $b(hT)^{-1}\le\sigma_h^2(x)\le bB(hT)^{-1}$ provide (4.1) with

$$h^*\asymp(TL^2D^2)^{-1/5}.$$

Hence, for the above-mentioned choice $\lambda_1\asymp\sqrt{\log\log T}$ and $D\asymp\lambda_1$, we obtain, due to Theorem 4.1, the following rate of the adaptive estimation

$$(\lambda+2\lambda_1+3D)\sigma_{h^*}(x)\asymp L^{1/5}\left(\frac{\log\log T}{T}\right)^{2/5}.$$

At the same time, the "ideal" choice of the bandwidth leads to the rate $L^{1/5}T^{-2/5}$, see Section 3.3. Thus, the adaptive rate is worse than the "ideal" one within a *log log*-factor only.

The origin of the $\log\log$-factor in the rate of adaptive estimation can be easily explained. The total number $\#\mathcal{H}$ of considered estimates is logarithmic in the observation time $T$ and the adaptive choice of the bandwidth leads to a worse accuracy by factor $\log(\#\mathcal{H})$ at some power.

The notion of "payment for adaptation" is now well understood in nonparametric estimation: if we have too many estimates to select between, we have to "pay" for the adaptive choice some additional factor in the risk of estimation. In particular, it is shown in Lepski (1990) and Brown and Low (1996) (see also Lepski and Spokoiny (1997)) that for the problem of pointwise adaptive estimation, the optimal adaptive rate has to be worse than the optimal one by a log-factor.

In our results a $\log\log$-factor appears. This fact is not in the contradiction with earlier issues, since the above-mentioned results correspond to the case of the power loss function $l(x) = |x|^p,\ p > 0$, whereas we consider the bounded loss function. It can be also shown that the rate achieved by our estimate is optimal for pointwise adaptive estimation with the bounded loss function (see Spokoiny (1997) for similar results in the adaptive testing problem).

# 5. Proofs

In this section we prove Theorems 3.1 and 4.1. For a generic positive constant the notation '$\ell$' will be used hereafter.

## 5.1. Decomposition of $\widetilde{f}_h(x)$

We use two obvious identities characterizing the local linear smoother: for $v_{1,h} = \frac{\mu_{1,h}}{D_h}$ and $v_{2,h} = \frac{\mu_{2,h}}{D_h}$

$$\begin{aligned}\int_0^T Q\left(\frac{X_s^\varepsilon - x}{h}\right)\left(v_{2,h} - v_{1,h}\frac{X_s^\varepsilon - x}{h}\right)\mathrm{d}s &= 1\\ \int_0^T Q\left(\frac{X_s^\varepsilon - x}{h}\right)\left(v_{2,h}\frac{X_s^\varepsilon - x}{h} - v_{1,h}\frac{(X_s^\varepsilon - x)^2}{h^2}\right)\mathrm{d}s &= 0\end{aligned}$$

and hence

$$\int_0^T Q\left(\frac{X_s^\varepsilon - x}{h}\right)\left(v_{2,h} - v_{1,h}\frac{X_s^\varepsilon - x}{h}\right)\overline{f}(x)\,\mathrm{d}s = \overline{f}(x) \tag{5.1}$$

$$\int_0^T Q\left(\frac{X_s^\varepsilon - x}{h}\right)\left(v_{2,h}\frac{X_s^\varepsilon - x}{h} - v_{1,h}\frac{(X_s^\varepsilon - x)^2}{h^2}\right)\overline{f}_x(x)\,\mathrm{d}s = 0. \tag{5.2}$$

Due to (2.3) and (1.1), the estimate $\widetilde{f}_h(x)$ can be represented as follows:

$$\begin{aligned}
\widetilde{f}_h(x) &= v_{2,h}\int_0^T Q\left(\frac{X^\varepsilon_s - x}{h}\right)\mathrm{d}X^\varepsilon_s - v_{1,h}\int_0^T Q\left(\frac{X^\varepsilon_s - x}{h}\right)\,\frac{X^\varepsilon_s - x}{h}\,\mathrm{d}X^\varepsilon_s \\
&= \int_0^T Q\left(\frac{X^\varepsilon_s - x}{h}\right)\left(v_{2,h} - v_{1,h}\frac{X^\varepsilon_s - x}{h}\right) f(X^\varepsilon_s, Y^\varepsilon_s)\,\mathrm{d}s \\
&\quad + v_{2,h}\int_0^T Q\left(\frac{X^\varepsilon_s - x}{h}\right) g(X^\varepsilon_s, Y^\varepsilon_s)\,\mathrm{d}w_s \\
&\quad - v_{1,h}\int_0^T Q\left(\frac{X^\varepsilon_s - x}{h}\right)\,\frac{X^\varepsilon_s - x}{h}\, g(X^\varepsilon_s, Y^\varepsilon_s)\,\mathrm{d}w_s .
\end{aligned}$$

Now (5.1) and (5.2) imply the following decomposition

$$\widetilde{f}_h(x) = \overline{f}(x) + \xi_h + r_h + \zeta^{(1)}_h + \zeta^{(2)}_h \tag{5.3}$$

where, with $\delta(X^\varepsilon_s, Y^\varepsilon_s, x) = f(X^\varepsilon_s, Y^\varepsilon_s) - f(x, Y^\varepsilon_s) - \dfrac{X^\varepsilon_s - x}{h}\, f_x(x, Y^\varepsilon_s)$,

$$\begin{aligned}
r_h &= \int_0^T Q\left(\frac{X^\varepsilon_s - x}{h}\right)\left(v_{2,h} - v_{1,h}\,\frac{X^\varepsilon_s - x}{h}\right)\delta(X^\varepsilon_s, Y^\varepsilon_s, x)\,\mathrm{d}s, \\
\xi_h &= v_{2,h}\int_0^T Q\left(\frac{X^\varepsilon_s - x}{h}\right) g(X^\varepsilon_s, Y^\varepsilon_s)\,\mathrm{d}w_s \\
&\quad - v_{1,h}\int_0^T Q\left(\frac{X^\varepsilon_s - x}{h}\right)\,\frac{X^\varepsilon_s - x}{h}\, g(X^\varepsilon_s, Y^\varepsilon_s)\,\mathrm{d}w_s, \\
\zeta^{(1)}_h &= v_{2,h}\int_0^T Q\left(\frac{X^\varepsilon_s - x}{h}\right)\left[f(x, Y^\varepsilon_s) - \overline{f}(x)\right]\mathrm{d}s \\
&\quad - v_{1,h}\int_0^T Q\left(\frac{X^\varepsilon_s - x}{h}\right)\left[f(x, Y^\varepsilon_s) - \overline{f}(x)\right]\frac{X^\varepsilon_s - x}{h}\,\mathrm{d}s, \\
\zeta^{(2)}_h &= v_{2,h}\int_0^T Q\left(\frac{X^\varepsilon_s - x}{h}\right)\left[f_x(x, Y^\varepsilon_s) - \overline{f}_x(x)\right]\frac{X^\varepsilon_s - x}{h}\,\mathrm{d}s \\
&\quad - v_{1,h}\int_0^T Q\left(\frac{X^\varepsilon_s - x}{h}\right)\left[f_x(x, Y^\varepsilon_s) - \overline{f}_x(x)\right]\frac{(X^\varepsilon_s - x)^2}{h^2}\,\mathrm{d}s.
\end{aligned}$$

Below we evaluate separately each term in this decomposition.

## 5.2. An upper bound for $|r_h|$

Since $Q\left(\frac{u-x}{h}\right)$ vanishes for any $u \notin [x-h, x+h]$ and $|\delta(X^\varepsilon_s, Y^\varepsilon_s, x)| \le \Delta_h(x)$ for $|X^\varepsilon_s - x| \le h$, we get

$$\begin{aligned}
|r_h| &\le \int_0^T Q\left(\frac{X^\varepsilon_s - x}{h}\right)\left|v_{2,h} - v_{1,h}\,\frac{X^\varepsilon_s - x}{h}\right|\,|\delta(X^\varepsilon_s, Y^\varepsilon_s, x)|\,\mathrm{d}s \\
&\le \Delta_h(x)\int_0^T Q\left(\frac{X^\varepsilon_s - x}{h}\right)\left|v_{2,h} - v_{1,h}\frac{X^\varepsilon_s - x}{h}\right|\mathrm{d}s.
\end{aligned} \tag{5.4}$$

The properties $|Q(u)| \le 1$ and $Q(u) = 0,\ |u| \ge 1$ imply the inequality $\mu_{2,h} \le \mu_{0,h}$. In addition we know that it holds on $\mathcal{A}_h$

$$\mu_{1,h}^2 \le \rho\mu_{0,h}\mu_{2,h}. \tag{5.5}$$

We now show that

$$|r_h| \le (1-\rho)^{-1/2}\Delta_h(x) \qquad \text{on} \quad \mathcal{A}_h. \tag{5.6}$$

The Cauchy-Schwarz inequality applied to (5.4) gives

$$|r_h| \le \Delta_h(x)\left\{\int_0^T Q\left(\frac{X_s^\varepsilon - x}{h}\right) \mathrm{d}s \int_0^T Q\left(\frac{X_s^\varepsilon - x}{h}\right)\left(v_{2,h} - v_{1,h}\frac{X_s^\varepsilon - x}{h}\right)^2 \mathrm{d}s\right\}^{1/2}.$$

Next,

$$\int_0^T Q\left(\frac{X_s^\varepsilon - x}{h}\right) \mathrm{d}s = \mu_{0,h},$$

and using $v_{k,h} = \mu_{k,h}/D_h$, with $D_h = \mu_{2,h}\mu_{0,h} - \mu_{1,h}^2$, $k = 0, 1, 2$, we get

$$\begin{aligned}
&\int_0^T Q\left(\frac{X_s^\varepsilon - x}{h}\right)\left(v_{2,h} - v_{1,h}\frac{X_s^\varepsilon - x}{h}\right)^2 \mathrm{d}s \\
&\quad = \frac{1}{D_h^2}\int_0^T Q\left(\frac{X_s^\varepsilon - x}{h}\right)\left(\mu_{2,h} - \mu_{1,h}\frac{X_s^\varepsilon - x}{h}\right)^2 \mathrm{d}s \\
&\quad = \frac{\mu_{2,h}^2}{D_h^2}\int_0^T Q\left(\frac{X_s^\varepsilon - x}{h}\right) \mathrm{d}s + \frac{\mu_{1,h}^2}{D_h^2}\int_0^T Q\left(\frac{X_s^\varepsilon - x}{h}\right)\frac{(X_s^\varepsilon - x)^2}{h^2} \mathrm{d}s \\
&\qquad - \frac{2\mu_{1,h}\mu_{2,h}}{D_h^2}\int_0^T Q\left(\frac{X_s^\varepsilon - x}{h}\right)\frac{X_s^\varepsilon - x}{h} \mathrm{d}s \\
&\quad = \frac{\mu_{2,h}^2\mu_{0,h} - \mu_{2,h}\mu_{1,h}^2}{D_h^2} \\
&\quad = \mu_{2,h}/D_h.
\end{aligned}$$

Hence, in view of (5.5),

$$|r_h| \ \le \ \Delta_h(x)\left(\frac{\mu_{0,h}\,\mu_{2,h}}{D_h}\right)^{1/2} = \Delta_h(x)\left(\frac{\mu_{0,h}\,\mu_{2,h}}{\mu_{0,h}\mu_{2,h} - \mu_{1,h}^2}\right)^{1/2} \le \Delta_h(x)\left(\frac{1}{1-\rho}\right)^{1/2}$$

as required.

## 5.3. An upper bound for $\boldsymbol{\xi_h}$

We study here some properties of the "stochastic term"

$$\begin{aligned}
\xi_h \ = \ & v_{2,h}\int_0^T Q\left(\frac{X_s^\varepsilon - x}{h}\right) g(X_s^\varepsilon, Y_s^\varepsilon)\,\mathrm{d}w_s \\
& - v_{1,h}\int_0^T Q\left(\frac{X_s^\varepsilon - x}{h}\right)\frac{X_s^\varepsilon - x}{h}\, g(X_s^\varepsilon, Y_s^\varepsilon)\,\mathrm{d}w_s.
\end{aligned}$$

Namely, we intend to show that the probability of the event $\{\xi_h > \lambda\sigma_h(x)\}$ with $\sigma_h(x)$ from (3.2) is small provided that $\lambda$ is large enough. Set for $t \le T$

$$
\begin{aligned}
M_{0,t} &= \int_0^t Q\left(\frac{X_s^\varepsilon - x}{h}\right) g(X_s^\varepsilon, Y_t^\varepsilon)\, \mathrm{d}w_s, \\
M_{1,t} &= \int_0^t Q\left(\frac{X_s^\varepsilon - x}{h}\right) \frac{X_s^\varepsilon - x}{h}\, g(X_s^\varepsilon, Y_t^\varepsilon)\, \mathrm{d}w_s.
\end{aligned}
$$

The Itô integrals $M_{0,t}$ and $M_{1,t}$ are continuous local martingales with the predictable quadratic variations (see e.g. Liptser and Shiryayev (1989))

$$
\begin{aligned}
\langle M_0\rangle_t &= \int_0^t Q^2\left(\frac{X_s^\varepsilon - x}{h}\right) g^2(X_s^\varepsilon, Y_s^\varepsilon)\, \mathrm{d}s, \\
\langle M_0, M_1\rangle_t &= \int_0^t Q^2\left(\frac{X_s^\varepsilon - x}{h}\right) \frac{X_s^\varepsilon - x}{h}\, g^2(X_s^\varepsilon, Y_s^\varepsilon)\, \mathrm{d}s, \\
\langle M_1\rangle_t &= \int_0^t Q^2\left(\frac{X_s^\varepsilon - x}{h}\right) \left(\frac{X_s^\varepsilon - x}{h}\right)^2 g^2(X_s^\varepsilon, Y_s^\varepsilon)\, \mathrm{d}s,
\end{aligned}
$$

so that $\langle M_0\rangle_T = V_{0,h}$, $\langle M_0, M_1\rangle_T = V_{1,h}$ and $\langle M_1\rangle_T = V_{2,h}$. This yields

$$
\begin{aligned}
\xi_h(x) &= v_{2,h} M_{0,T} - v_{1,h} M_{1,T}, \\
\sigma_h^2(x) &= v_{2,h}^2 \langle M_0\rangle_T - 2 v_{1,h} v_{2,h} \langle M_0, M_1\rangle_T + v_{1,h}^2 \langle M_1\rangle_T.
\end{aligned}
$$

Denote

$$
u_h = \frac{v_{1,h}}{v_{2,h}} = \frac{\mu_{1,h}}{\mu_{2,h}}.
$$

Obviously

$$
\begin{aligned}
&\boldsymbol{P}\left(|\xi_h| > \lambda \sigma_h(x), \mathcal{A}_h\right) \\
&\quad = \boldsymbol{P}\left(|M_{0,T} - u_h M_{1,T}| > \lambda \sqrt{\langle M_0\rangle_T - 2u_h \langle M_0, M_1\rangle_T + u_h^2 \langle M_1\rangle_T},\ \mathcal{A}_h\right).
\end{aligned}
$$

To evaluate the latter probability, we apply the general result from Proposition 6.2, see Appendix. First we check the required conditions. The value $u_h$, being restricted to $\mathcal{A}_h$, can be bounded as:

$$
|u_h| \le \left|\frac{\sqrt{\rho \mu_{0,h} \mu_{2,h}}}{\mu_{2,h}}\right| \le \sqrt{\rho r}.
$$

Note now that

$$
\begin{aligned}
\frac{\langle M_1\rangle_T}{\langle M_0\rangle_T - 2u_h \langle M_0, M_1\rangle_T + u_h^2 \langle M_1\rangle_T} &= \frac{V_{2,h}}{V_{0,h} - 2u_h V_{1,h} + u_h^2 V_{2,h}} \\
&= \frac{V_{2,h}^2}{V_{0,h} V_{2,h} - V_{1,h}^2 + (V_{1,h} - u_h V_{2,h})^2},
\end{aligned}
$$

and it holds on $\mathcal{A}_h$ in view of $V_{2,h} \le V_{0,h}$

$$
\frac{\langle M_1\rangle_T}{\langle M_0\rangle_T - 2u_h \langle M_0, M_1\rangle_T + u_h^2 \langle M_1\rangle_T} \le \frac{V_{2,h}^2}{(1-\rho) V_{0,h} V_{2,h}} \le \frac{1}{1-\rho}.
$$

In addition, the definition of $\mathcal{A}_h$ provides the following bounds for $\sigma_h^2(x)$ on this set

$$\frac{\sigma_h^2(x)}{Th\,v_{2,h}^2} = \frac{Th\,\sigma_h^2(x)}{(Th\,v_{2,h})^2} \le \frac{bB}{b^2} = \frac{B}{b},$$
$$\frac{\sigma_h^2(x)}{Th\,v_{2,h}^2} = \frac{Th\,\sigma_h^2(x)}{(Th\,v_{2,h})^2} \ge \frac{b}{(bB)^2} = \frac{1}{bB^2}.$$

Applying now Proposition 6.2 we get

$$\boldsymbol{P}\left(|\xi_h| > \lambda\sigma_h(x), \mathcal{A}_h\right) \le 4e\log(4B^3)\left(1 + 4r\sqrt{\frac{1+r}{1-\rho}}\,\lambda^2\right)\lambda e^{-\frac{\lambda^2}{2}}. \tag{5.7}$$

## 5.4. An upper bounds for $\zeta_h^{(1)}$ and $\zeta_h^{(2)}$

Note that both $\zeta_h^{(1)}$, $\zeta_h^{(2)}$ are linear combinations of elements of the form

$$v_h \int_0^T \Psi(X_s^\varepsilon)[a(Y_s^\varepsilon) - \overline{a}]\,\mathrm{d}s,$$

where

- $v_h$ is any of $v_{1,h}$, $v_{2,h}$;
- $\Psi(X_s^\varepsilon)$ is any of $\frac{(X_s^\varepsilon - x)^k}{h^k} Q\left(\frac{X_s^\varepsilon - x}{h}\right)$, $k = 0, 1, 2$;
- $a(Y_s^\varepsilon)$ is any of $f(x, Y_s^\varepsilon)$, $f_x(x, Y_s^\varepsilon)$, and $\overline{a} = \int a(y)\,p(y)\,\mathrm{d}y$, with $p(\cdot)$ being the invariant density of the fast process.

Under the assumptions made, the function $\Psi(u)$ is bounded by 1 and twice continuously differentiable: there exists a constant $C_1$ such that

$$|\Psi(u)| \le 1 \text{ and } |\dot\Psi(u)| + |\ddot\Psi(u)| \le C_1 \qquad \forall u.$$

Next, on the set $\mathcal{A}_h$ it holds $v_{1,h}^2 \le \rho v_{0,h} v_{2,h} \le \rho r v_{2,h}^2$ and $v_{2,h} \le bB(Th)^{-1}$, so that, taking into account $Th \ge 1$, it suffices to bound only

$$U_T^\varepsilon = \int_0^T \Psi(X_s^\varepsilon)[a(Y_s^\varepsilon) - \overline{a}]\,\mathrm{d}s.$$

We apply a large deviation type estimate for the two-scaled diffusion model (1.1), (1.2) from Liptser and Spokoiny (1997) adapted to the case considered.

**Proposition 5.1.** *Suppose* $(A_s)$ *and* $(A_f)$. *If* $T = T_\varepsilon$ *and* $\lim_{\varepsilon\to 0} \varepsilon T_\varepsilon = 0$, *then for every positive* $z > 0$ *and* $0 < \kappa < 1/2$

$$\varlimsup_{\varepsilon\to 0}(\varepsilon T_\varepsilon)^{1-2\kappa}\log \boldsymbol{P}\left((\varepsilon T_\varepsilon)^{-\kappa}|U_{T_\varepsilon}^\varepsilon| > z\right) \le -\frac{z^2}{2\gamma},$$

*where*

$$\gamma = \int_{\boldsymbol{R}} \vartheta^2(y)\,G^2(y)\,p(y)\,\mathrm{d}y,$$
$$\vartheta(y) = \frac{2}{G^2(y)\,p(y)} \int_\infty^y [a(u) - \overline{a}]\,p(u)\,\mathrm{d}u.$$

**Corollary 5.1.** *For $\varepsilon$ small enough and $\kappa_1 < 1-2\kappa$*

$$\boldsymbol{P}\left(|U^{\varepsilon}_{T^{\varepsilon}}| > (\varepsilon T^{\varepsilon})^{\kappa}\right) < \exp\left(-\frac{1}{(\varepsilon T^{\varepsilon})^{\kappa_1}}\right).$$

Applying this corollary with $\kappa < 1/2$ and $\kappa_1 < 1-2\kappa$, we obtain for $\varepsilon T$ small enough

$$\boldsymbol{P}\left(|\zeta^{(i)}_h| > 2(\varepsilon T)^{\kappa}\right) < 2\exp\left(-\frac{1}{(\varepsilon T)^{\kappa_1}}\right), \qquad i = 1,2. \tag{5.8}$$

## 5.5. Proof of Theorem 3.1

Summing up the decomposition (5.3) and the bounds (5.6), (5.7), (5.8), we get

$$\begin{aligned}
&\boldsymbol{P}\left(\left|\widetilde{f}_h(x) - \overline{f}(x)\right| > c\Delta_h(x) + \lambda\sigma_h(x) + 2(\varepsilon T^{\varepsilon})^{\kappa},\, \mathcal{A}_h\right) \\
&\quad \le 4e\log(4B^3)\left(1 + 4r\sqrt{\frac{1+r}{1-\rho}}\,\lambda^2\right)\lambda e^{-\frac{\lambda^2}{2}} + 4\exp\left(-\frac{1}{(\varepsilon T)^{\kappa_1}}\right).
\end{aligned}$$

This leads to the required bound from Theorem 3.1 for sufficiently small $\varepsilon T$ .

## 5.6. Proof of Theorem 4.1

Let $h^*$ be shown in the theorem. Recall that $\mathcal{A}^* = \bigcap_{h\in\mathcal{H}} \mathcal{A}_h$. We use an obvious inequality

$$\begin{aligned}
&\boldsymbol{P}\left(\left|\widehat{f}(x) - \overline{f}(x)\right| > (\lambda + 2\lambda_1 + 3D)\sigma_{h^*}(x),\, \mathcal{A}^*\right) \\
\le\ &\boldsymbol{P}\left(\left|\widehat{f}(x) - \overline{f}(x)\right| > (\lambda + 2\lambda_1 + 3D)\sigma_{h^*}(x),\, \widehat{h} \ge h^*,\, \mathcal{A}^*\right) + \boldsymbol{P}\left(\widehat{h} < h^*,\, \mathcal{A}^*\right).
\end{aligned}$$

Since $\sigma_h(x)$ decreases in $h$, we have on the set $\{\widehat{h} \ge h^*\} \cap \mathcal{A}^*$ in view of the definition of $\widehat{h}$

$$|\widetilde{f}_{\widehat{h}}(x) - \widetilde{f}_{h^*}(x)| \le \lambda_1\big(\sigma_{\widehat{h}}(x) + \sigma_{h^*}(x)\big) + 2D\sigma_{\widehat{h}}(x) \le 2(\lambda_1 + D)\sigma_{h^*}(x).$$

Further, using the inequality $c\Delta_{h^*}(x) \le D\sigma_{h^*}(x)$ and Theorem 3.1, we get

$$\begin{aligned}
&\boldsymbol{P}\left(|\widetilde{f}_{h^*}(x) - \overline{f}(x)| > (D+\lambda)\sigma_{h^*}(x), \mathcal{A}^*\right) \\
&\quad \le \boldsymbol{P}\left(|\widetilde{f}_{h^*}(x) - \overline{f}(x)| > \lambda\sigma_{h^*}(x) + c\Delta_{h^*}(x)\,,\, \mathcal{A}^*\right) \\
&\quad \le \left(C_1\lambda + C_2\lambda^3\right) e^{-\frac{\lambda^2}{2}},
\end{aligned}$$

where

$$\begin{aligned}
C_1 &= 4e\log(4B^3), \\
C_2 &= 4e\log(4B^3)\,4r\sqrt{\frac{1+r}{1-\rho}}\,.
\end{aligned}$$

Hence

$$\boldsymbol{P}\Big(|\widehat{f}(x)-\overline{f}(x)|>(\lambda+2\lambda_1+3D)\sigma_{h^*}(x),\,\mathcal{A}^*,\widehat{h}\ge h^*\Big) \le \left(C_1\lambda+C_2\lambda^3\right)e^{-\frac{\lambda^2}{2}} \tag{5.9}$$

and it only remains to evaluate $\boldsymbol{P}(\widehat{h}<h^*,\,\mathcal{A}^*)$. Due to the definition of $\widehat{h}$, we have

$$\{\widehat{h}<h^*,\,\mathcal{A}^*\} \subseteq \bigcup_{h\in\mathcal{H}\,:\,h<h^*}\;\bigcup_{\eta\in\mathcal{H}\,:\,\eta<h}\Big\{|\widehat{f}_h(x)-\widehat{f}_\eta(x)|>\lambda_1\Big(\sigma_h(x)+\sigma_\eta(x)\Big)+2D\sigma_h(x),\,\mathcal{A}^*\Big\}.$$

We now use that for every $\eta,h\in\mathcal{H}$ with $\eta<h<h^*$

$$\begin{aligned} c\Delta_h(x) &\le c\Delta_{h^*}(x)\le D\sigma_{h^*}(x)\le D\sigma_h(x),\\ c\Delta_\eta(x) &\le c\Delta_{h^*}(x)\le D\sigma_{h^*}(x)\le D\sigma_h(x). \end{aligned}$$

Therefore by Theorem 3.1

$$\begin{aligned} &\boldsymbol{P}\Big(|\widetilde{f}_h(x)-\widetilde{f}_\eta(x)|>\lambda_1\big(\sigma_h(x)+\sigma_\eta(x)\big)+2D\sigma_h(x)\,,\,\mathcal{A}^*\Big)\\ &\le \boldsymbol{P}\Big(|\widetilde{f}_h(x)-\overline{f}(x)|>\lambda_1\sigma_h(x)+c\Delta_h(x)\,,\,\mathcal{A}_h\Big)\\ &\quad+\boldsymbol{P}\Big(|\widetilde{f}_\eta(x)-\overline{f}(x)|>\lambda_1\sigma_\eta(x)+c\Delta_\eta(x)\,,\,\mathcal{A}_\eta\Big)\\ &\le 2\left(C_1\lambda_1+C_2\lambda_1^3\right)e^{-\frac{\lambda_1^2}{2}}. \end{aligned}$$

Clearly the total number of pairs $\eta,h\in\mathcal{H}$, satisfying $\eta<h<h^*$, is at most $(\#\mathcal{H})^2/2$. Therefore

$$\boldsymbol{P}\left(\widehat{h}<h^*\right)\le(\#\mathcal{H})^2\left(C_1\lambda_1+C_2\lambda_1^3\right)e^{-\frac{\lambda_1^2}{2}}.$$

This bound coupled with (5.9) implies the desired assertion.

# 6. Appendix. Deviation probabilities for martingales

In the Appendix we present two general results for continuous martingales. The first result describes some properties of real-valued martingales, while the second one deals with martingales valued in $\mathbb{R}^2$.

## 6.1. The scalar case

Let $M_t$ be a continuous martingale with $M_0=0$ and with the predictable quadratic variation $\langle M\rangle_t$.

**Proposition 6.1.** *For every* $T>0$, $\vartheta>0$, $S\ge1$ *and* $\lambda\ge1$

$$\boldsymbol{P}\left(|M_T|>\lambda\sqrt{\langle M\rangle_T},\ \vartheta\le\sqrt{\langle M\rangle_T}\le\vartheta S\right)\le 4\lambda\sqrt{e}\,(1+\log S)\,e^{-\frac{\lambda^2}{2}}.$$

*Proof.* We use

$$\begin{aligned} &\boldsymbol{P}\left(|M_T| > \lambda\sqrt{\langle M\rangle_T},\ \vartheta \le \sqrt{\langle M\rangle_T} \le \vartheta S\right) \\ &\quad \le \boldsymbol{P}\left(M_T > \lambda\sqrt{\langle M\rangle_T},\ \vartheta \le \sqrt{\langle M\rangle_T} \le \vartheta S\right) \\ &\qquad + \boldsymbol{P}\left(M_T < -\lambda\sqrt{\langle M\rangle_T},\ \vartheta \le \sqrt{\langle M\rangle_T} \le \vartheta S\right). \end{aligned}$$

We estimate separately each term in the right side of this inequality.

Given $a > 1$, introduce the geometric series $\vartheta_k = \vartheta a^k$ and define the sequence of random events $\mathcal{C}_k = \{\vartheta_k \le \sqrt{\langle M\rangle_T} < \vartheta_{k+1}\}$, $k = 0, 1, \dots$. Then clearly

$$\begin{aligned} &\boldsymbol{P}\left(M_T > \lambda\sqrt{\langle M\rangle_T},\ \vartheta \le \sqrt{\langle M\rangle_T} \le \vartheta S\right) \\ &\quad \le \sum_{k\ge 0}^{K} \boldsymbol{P}\left(M_T > \lambda\sqrt{\langle M\rangle_T},\ \vartheta \le \sqrt{\langle M\rangle_T} \le \vartheta S,\ \mathcal{C}_k\right). \end{aligned} \tag{6.1}$$

where $K$ is the integer part of $\log_a S$. We now bound each term in this sum. Let, with $\gamma \in \mathbb{R}$,

$$Z_t(\gamma) = \exp\left(\gamma M_t - \frac{\gamma^2}{2}\langle M\rangle_t\right).$$

The random process $Z_t(\gamma)$ is the continuous local martingale and, being positive, it is the supermartingale (see Problem 1.4.4 in Liptser and Shiryayev (1986)). Therefore for every $T > 0$,

$$\boldsymbol{E} Z_T(\gamma) \le 1. \tag{6.2}$$

For fixed $k$, we pick $\gamma_k = \frac{\lambda}{\vartheta_k}$ and use (6.2) for the inequality

$$1 \ge \boldsymbol{E} Z_T(\gamma_k) \boldsymbol{I}\left(M_T > \lambda\sqrt{\langle M\rangle_T},\ \mathcal{C}_k\right)$$

which implies

$$\begin{aligned} 1 &\ge \boldsymbol{E}\exp\left(\frac{\lambda}{\vartheta_k} M_T - \frac{\lambda^2}{2\vartheta_k}\langle M\rangle_T\right)\boldsymbol{I}\left(M_T > \lambda\sqrt{\langle M\rangle_T}, \mathcal{C}_k\right) \\ &\ge \boldsymbol{E}\exp\left(\frac{\lambda^2}{\vartheta_k}\sqrt{\langle M\rangle_T} - \frac{\lambda^2}{2\vartheta_k}\langle M\rangle_T\right)\boldsymbol{I}\left(M_T > \lambda\sqrt{\langle M\rangle_T}, \mathcal{C}_k\right) \\ &\ge \boldsymbol{E}\exp\left\{\inf_{\vartheta_k \le v \le \vartheta_{k+1}}\left(\frac{\lambda^2 v}{\vartheta_k} - \frac{\lambda^2 v^2}{2\vartheta_k^2}\right)\right\}\boldsymbol{I}\left(M_T > \lambda\sqrt{\langle M\rangle_T}, \mathcal{C}_k\right). \end{aligned}$$

It is easy to check that "$\inf_{\vartheta_k \le v \le \vartheta_{k+1}}$" is attained at the point $v = \vartheta_{k+1} = a\vartheta_k$ so that

$$\boldsymbol{P}\left(M_T > \lambda\sqrt{\langle M\rangle_T}, \mathcal{C}_k\right) \le \exp\left\{-\lambda^2\left(a - \frac{a^2}{2}\right)\right\}.$$

Combining this bound with (6.1) and using $K \le \log_a S$, we obtain

$$\boldsymbol{P}\left(M_T > \lambda\sqrt{\langle M\rangle_T}, \vartheta \le \sqrt{\langle M\rangle_T} \le \vartheta S\right) \le (1 + \log_a S)\exp\left\{-\lambda^2\left(a - \frac{a^2}{2}\right)\right\}.$$

Since the left hand side of this inequality does not depend on $a$, we may optimize the choice of $a$ to minimize its right side. This leads to $a = 1+1/\lambda$ and $\lambda^2\left(a-\frac{a^2}{2}\right) = \lambda^2\left\{1+\frac{1}{\lambda}-\frac{1}{2}\left(1+\frac{1}{\lambda}\right)^2\right\} = \frac{1}{2}(\lambda^2-1)$. Since also $\log(1+1/\lambda) \geq 1/(2\lambda)$ for $\lambda \geq 1$, and hence $\log_a S \leq 2\lambda \log S$, we have

$$\boldsymbol{P}\left(M_T > \lambda\sqrt{\langle M\rangle_T},\, \vartheta \leq \sqrt{\langle M\rangle_T} \leq \vartheta S\right) \leq 2\sqrt{e}\lambda\,(1+\log S)\,e^{-\frac{\lambda^2}{2}}.$$

In the similar way we obtain

$$\boldsymbol{P}\left(M_T < -\lambda\sqrt{\langle M\rangle_T},\, \vartheta \leq \sqrt{\langle M\rangle_T} \leq \vartheta S\right) \leq 2\sqrt{e}\lambda\,(1+\log S)\,e^{-\frac{\lambda^2}{2}}$$

and the assertion follows. □

## 6.2. The vector case

Here, we consider continuous vector martingale $M_t$ valued in $\mathbb{R}^2$ with components $M_{0,t}$ and $M_{1,t}$, $t \geq 0$. We denote

$$\begin{aligned} V_{0,t} &= \langle M_0\rangle_t\,, \\ V_{1,t} &= \langle M_0, M_1\rangle_t\,, \\ V_{2,t} &= \langle M_1\rangle_t\,. \end{aligned}$$

Let $u$ be a random variable and

$$\sigma_t^2 = V_{0,t} - 2uV_{1,t} + u^2 V_{2,t}.$$

For a fixed time moment $T$ and constants $\vartheta > 0$, $S \geq 1$, $\beta \geq 0$ and $\rho \in (0,1)$, introduce the event

$$\mathcal{A}_T = \left\{\begin{array}{c} \vartheta \leq \sigma_T^2 \leq \vartheta S \\ V_{1,T}^2 \leq \rho V_{0,T} V_{2,T} \\ |u| \leq \beta \end{array}\right\}. \tag{6.3}$$

**Proposition 6.2.** *Let $M_t$ be a martingale with values in $\mathbb{R}^2$ such that $V_{0,T} \geq V_{2,T}$. Then, with $\mathcal{A}_T$ from (6.3), it holds for every $\lambda \geq \sqrt{2}$,*

$$\boldsymbol{P}\left(|M_{0,T} - uM_{1,T}| > \lambda\sigma_T,\, \mathcal{A}_T\right) \leq 4e\log(4S)\left(1 + 4\beta\sqrt{\frac{1+\beta}{1-\rho}}\,\lambda^2\right)\lambda e^{-\frac{\lambda^2}{2}}.$$

*Proof.* For fixed $\beta$, $\rho$, and $\lambda$ define $\delta$ such that

$$\frac{2\delta(1+\beta)}{1-\rho} = \lambda^{-2} \tag{6.4}$$

and denote by $D_\delta = \{\alpha_k = k\delta : k \in \mathbb{N},\, |\alpha| \leq \beta\}$ the discrete grid with the step $\delta$ in the interval $[-\beta, \beta]$.

Let $\nu_+$ (respectively $\nu_-$) be random variable valued in $D_\delta$ which is closest to $u$ from above (respectively from below). Then clearly

$$|\nu_\pm - u| \leq \delta. \tag{6.5}$$

$$|M_{0,T} - uM_{1,T}| \leq \max\left\{|M_{0,T} - \nu_- M_{1,T}|\,, |M_{0,T} - \nu_+ M_{1,T}|\right\}. \tag{6.6}$$

Let now $\nu$ be one of $\nu_-$ and $\nu_+$. Then by construction $|\nu - u| \le \delta$. Next we show that on the set $\mathcal{A}_T$ it holds

$$1 - \lambda^{-2} \le \frac{V_{0,T} - 2\nu V_{1,T} + \nu^2 V_{2,T}}{\sigma_T^2} \le 1 + \lambda^{-2} \tag{6.7}$$

Indeed

$$\begin{aligned}
\sigma_T^2 &= V_{0,T} - 2uV_{1,T} + u^2 V_{2,T} \\
&= V_{0,T} - \frac{V_{1,T}^2}{V_{2,T}} + V_{2,T}\left(u - \frac{V_{1,T}}{V_{2,T}}\right)^2 \\
&\ge \frac{V_{0,T}V_{2,T} - V_{1,T}^2}{V_{2,T}} \\
&\ge (1-\rho)V_{0,T}
\end{aligned}$$

and using $V_{2,T} \le V_{0,T}$, we get

$$\begin{aligned}
\frac{|V_{1,T}|}{\sigma_T^2} &\le \frac{\sqrt{\rho V_{0,T} V_{2,T}}}{(1-\rho)V_{0,T}} \le \frac{\sqrt{\rho}}{1-\rho} \le (1-\rho)^{-1}, \\
\frac{V_{2,T}}{\sigma_T^2} &\le \frac{V_{2,T}}{(1-\rho)V_{0,T}} \le (1-\rho)^{-1}.
\end{aligned}$$

Since on the set $\mathcal{A}$ it holds $|u| \le \beta$ and by construction $|\nu| \le \beta$, we obtain, using definition (6.4) of $\delta$,

$$\begin{aligned}
&\left|V_{0,T} - 2uV_{1,T} + u^2V_{2,T} - (V_{0,T} - 2\nu V_{1,T} + \nu^2 V_{2,T})\right| \\
&\quad \le 2|V_{1,T}||u-\nu| + V_{2,T}\left|u^2 - \nu^2\right| \\
&\quad \le 2\delta(1-\rho)^{-1}\sigma_T^2 + 2\beta\delta(1-\rho)^{-1}\sigma_T^2 \\
&\quad = \sigma_T^2 \lambda^{-2}
\end{aligned}$$

and (6.7) follows.

Since on the set $\mathcal{A}_T$ the value $\sigma_T^2$ is between $\vartheta$ and $\vartheta S$, we also get for $\nu = \nu_\pm$

$$(1-\lambda^{-2})\vartheta \le V_{0,T} - 2\nu V_{1,T} + \nu^2 V_{2,T} \le (1+\lambda^{-2})\vartheta S. \tag{6.8}$$

We now derive from (6.6), (6.7) and (6.8)

$$\begin{aligned}
&\{M_{0,T} - uM_{1,T}| > \lambda\sigma_T,\, \mathcal{A}_T\} \\
&\quad \subseteq \left\{M_{0,T} - \nu_- M_{1,T}| > \frac{\lambda}{\sqrt{1+\lambda^2}}\sqrt{V_{0,T} - 2\nu_- V_{1,T} + \nu_-^2 V_{2,T}}\ ,\, \mathcal{A}_T\right\} \\
&\qquad \cup \left\{M_{0,T} - \nu_+ M_{1,T}| > \frac{\lambda}{\sqrt{1+\lambda^2}}\sqrt{V_{0,T} - 2\nu_+ V_{1,T} + \nu_+^2 V_{2,T}}\ ,\, \mathcal{A}_T\right\} \\
&\quad \subseteq \bigcup_{\alpha \in D_\delta} \left\{|M_{0,T} - \alpha M_{1,T}| > \frac{\lambda}{\sqrt{1+\lambda^2}}\sqrt{V_{0,T} - 2\alpha V_{1,T} + \alpha^2 V_{2,T}}\,,\, \mathcal{A}_{\alpha,T}\right\},
\end{aligned}$$

where

$$A_{\alpha,T} = \left\{(1-\lambda^{-2})\vartheta \le V_{0,T} - 2\alpha V_{1,T} + \alpha^2 V_{2,T} \le (1+\lambda^{-2})\vartheta S\right\}.$$

Now, for every $\alpha \in D_\delta$ , the process $M_{0,t}-\alpha M_{1,t}$ is the continuous local martingale with $\langle M_0 - \alpha M_1\rangle_T = V_{0,T} - 2\alpha V_{1,T} + \alpha^2 V_{2,T}$ . Applying Proposition 6.1 and using the inequalities $\lambda^2 \ge 2$ and $\frac{\lambda^2}{1+\lambda^{-2}} \ge \lambda^2(1-\lambda^{-2}) = \lambda^2 - 1$, we obtain

$$\begin{aligned}
&\boldsymbol{P}\left(|M_{0,T} - \alpha M_{1,T}| > \frac{\lambda}{\sqrt{1+\lambda^2}}\sqrt{V_{0,T} - 2\alpha V_{1,T} + \alpha^2 V_{2,T}}\,,\, A_{\alpha,T}\right) \\
&\quad \le 4\frac{\lambda}{\sqrt{1+\lambda^{-2}}}\left(1 + \log\frac{(1+\lambda^{-2})\vartheta S}{(1-\lambda^{-2})\vartheta}\right)\exp\left(-\frac{\lambda^2}{2(1+\lambda^{-2})} + \frac{1}{2}\right) \\
&\quad \le 4\lambda\left(1 + \log\frac{3S}{2}\right)\exp\left(-\frac{\lambda^2}{2} + 1\right).
\end{aligned}$$

Since the number of different elements in $D_\delta$ is at most $1 + 2\beta\delta^{-1}$ and since $\delta$ from (6.4) fulfills $\delta^{-1} = \frac{2\lambda^2(1+\beta)}{1-\rho}$ we get

$$\begin{aligned}
\boldsymbol{P}\left(|M_{0,T} - uM_{1,T}| > \lambda\sigma_T\,,\, \mathcal{A}_T\right) &\le 4e\left(1 + \log\frac{3S}{2}\right)\left(1 + 2\beta\delta^{-1}\right)\lambda e^{-\frac{\lambda^2}{2}} \\
&\le 4e\log(4S)\left(1 + 4\beta\sqrt{\frac{1+\beta}{1-\rho}}\,\lambda^2\right)\lambda e^{-\frac{\lambda^2}{2}}
\end{aligned}$$

as required. □

# References


[1] Brown, L.D. and Low, M.G. (1996). A constrained risk inequality with applications to nonparametric functional estimation. *Ann.-Statist.* **24** (1996), no. 6, 2524–2535.

[2] Carrol, R.J. and Hall, P. (1988). Optimal rate of convergence for deconvoluting a density. *J. Amer. Statist. Assoc.* **83**, 1184–1186.

[3] G. Collomb and P. Doukhan (1983). Estimation non parametrique de la fonction d'autoregression d'un processus stationnaire et phi melangeant: risques quadratiques pour la methode du noyau, *C. R. Acad. Sci.*, Paris, Ser. I, **296**, 859-862 .

[4] Delyon, B. and Juditsky, A. (1997). On minimax prediction for nonparametric autoregressive models. Unpublished manuscript.

[5] P. Doukhan and M. Ghindes (1980). Estimations dans le processus "$X_{n+1} = f(X_n) + epsilon_n$", *C. R. Acad. Sci.*, Paris, Ser. A **291**, 61-64.

[6] Doukhan, P. and Tsybakov, A.B. (1993). Nonparametric recurrent estimation in nonlinear ARX models. *Problemy-Peredachi-Informatsii* **29**, no. 4, 24–34. Translation: *Problems Inform. Trans.* **29**, no. 4, 318–327.

[7] Fan, J. and Truong, Y.K. (1993). Nonparametric regression with errors in variables. *Ann. Statist.* **21**, 1900–1925.

[8] Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications.* Chapman and Hall, London.

[9] Freidlin, M.I., Wentzell A.D. (1984). *Random Perturbations of Dynamical Systems.* Springer. N.Y

[10] Genon-Catalot, V., Laredo, C. and Picard, D. (1992). Nonparametric estimation of the diffusion coefficient by wavelet methods, *Scand. J. Statist.* **19**, 317-335.

[11] W. Häerdle and P. Vieu (1992). Kernel regression smoothing of time series", *J. Time Ser. Anal.* **13**, No.3, 209-232.

[12] Grama, I. and Nussbaum, M. (1998). Asymptotic equivalence for nonparametric generalized linear models, *Prob. Theory and Rel. Fields,* **111**, 167–214.

[13] Ibragimov,I.A. and Khasminskii,R.Z. (1981). *Statistical Estimation: Asymptotic Theory* Springer, New York.

[14] Katkovnik, V. Ja. (1985). *Nonparametric Identification and Data Smoothing: Local Approximation Approach*. Nauka, Moscow (in Russian).

[15] Khasminskii, R.Z. (1966). On stochastic processes defined by differential equations with small parameter. *Theory Probab. Appl.* **11**, 219–228.

[16] H.J.Kushner (1990). *Weak Convergence Methods and Singularly Perturbed Stochastic Control and Filtering Problems*. Birkhäuser 1990.

[17] Kutoyants, Yu.A. (1984a). On nonparametric estimation of trend coefficients in a diffusion process. Collection: *Statistics and control of stochastic processes*, Moscow, 230–250.

[18] Kutoyants, Yu.A. (1984b). Parameter estimation for stochastic processes. Translated from the Russian and edited by B. L. S. Prakasa Rao. *R & E Research and Exposition in Mathematics*, **6**. Heldermann Verlag, Berlin.

[19] Lepski, O. (1990). One problem of adaptive estimation in Gaussian white noise. *Theory Probab. Appl.* **35**, no. 3, 459–470.

[20] Lepski, O. and Levit, B. (1997). Efficient adaptive estimation of infinitely differentiable function. *Math. Methods of Statistics*, submitted.

[21] Lepski, O., Mammen, E. and Spokoiny, V. (1997). Ideal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selection. *Annals of Statistics*, **25**, no.3, 929–947.

[22] Lepski, O. and Spokoiny, V. (1997). Optimal pointwise adaptive methods in nonparametric estimation. *Annals of Statistics*, **25**, no.6,

[23] Liptser, R. and Shiryaev, A. (1989). *Theory of Martingales*. Kluwer Acad. Publ. 1989.

[24] R.S. Liptser, W.J. Runggaldier and M. Taksar (1996). Deterministic approximation for stochastic control problem. *SIAM J. on Control and Optimization.* **34** 161–178.

[25] Liptser, R. and Spokoiny, V. (1997). Moderate deviations for integral functionals of diffusion process. Unpublished manuscript.

[26] Spokoiny, V. (1996). Adaptive hypothesis testing using wavelets. *Annals of Stat.*, **24**, no. 6, 2477–2498.

[27] Milstein, G. and Nussbaum, M. (1994). Nonparametric estimation of a nonparametric diffusion model, *Prob. Theory and Rel. Fields.* To appear.

[28] Tsybakov, A. (1986). Robust reconstruction of functions by the local approximation. *Prob. Inf. Transm.*, **22**, 133-146.

[29] Veretennikov, A. Yu. (1991) On the averaging principle for systems of stochastic differential equations. *Math. USSR Sborn.*, **69**, No. 1, 271-284.

[30] Veretennikov, A. Yu. (1992) On large deviations for ergodic empirical measures. *Topics in Nonparametric Estimation. Advances in Soviet Mathematics*, AMS **12**, 125-133.



Dept. Electrical Engineering-Systems, Tel Aviv University, 69978 Tel Aviv, Israel

*E-mail address*: liptser@eng.tau.ac.il

Weierstrass Institute for Applied Analysis and Stochastics, Mohrenstr. 39, 10117 Berlin, Germany

*E-mail address*: spokoiny@wias-berlin.de